Nucleotides are ubiquitous molecules with considerable structural diversity. There are eight common varieties of nucleotides, each composed of a nitrogenous base linked to a sugar to which at least one phosphate group is also attached. The bases of nucleotides are planar, aromatic, heterocyclic molecules that are struc¬ tural derivatives of either purine or pyrimidine (although they are not syn¬ thesized in vivo from either of these organic compounds).

The most common purines are adenine (A) and guanine (G), and the major pyrimidines are cytosine (C), uracil (U), and thymine (T). The purines form bonds to a five-carbon sugar (a pentose) via their N9 atoms, whereas pyrimidines do so through their N1 atoms (Table 3-1). In ribonucleotides, the pentose is ribose, while in deoxyribonucleotides (or just deoxynucleotides), the sugar is 2'-deoxyribose (i.e., the carbon at position 2' lacks a hydroxyl group).

Note that the “primed” numbers refer to the atoms of the pentose; “unprimed” numbers refer to the atoms of the nitrogenous base.

In a ribonucleotide or a deoxyribonucleotide, one or more phosphate groups are bonded to atom C3' or atom C5' of the pentose to form a 3'-nucleotide or a 5'-nucleotide, respectively (Fig. 3-1). When the phosphate group is absent, the compound is known as a nucleoside. A 5'-nucleotide can therefore be called a nucleoside-5'-phosphate. Nucleotides most commonly contain one to three phosphate groups at the C5' position and are called nu¬ cleoside monophosphates, diphosphates, and triphosphates.

The structures, names, and abbreviations of the common bases, nucleo¬ sides, and nucleotides are given in Table 3-1. Ribonucleotides are components of RNA (ribonucleic acid), whereas deoxynucleotides are components of DNA (deoxyribonucleic acid). Adenine, guanine, and cytosine occur in both ribonucleotides and deoxynucleotides (accounting for six of the eight common nucleotides), but uracil primarily occurs in ribonucleotides and thymine oc¬ curs in deoxynucleotides. Free

nucleotides, which are anionic, are almost al¬ ways associated with the counterion Mg 2+ in cells. Nucleotides Participate in Metabolic Reactions. The bulk of the nucleotides in any cell are found in polymeric forms, as either DNA or RNA, whose pri¬ mary functions are information storage and transfer. However, free nucleotides and nucleotide derivatives perform an enormous variety of metabolic functions Section 1 Nucleotides not related to the management of genetic information. Perhaps the best known nucleotide is adenosine triphosphate (ATP), a nucleotide containing adenine, ribose, and a triphosphate group. ATP is often mistakenly referred to as an energy-storage molecule, but it is more accurately termed an energy carrier or energy transfer agent. The process of photosynthe¬ sis or the breakdown of metabolic fuels such as carbohydrates and fatty acids leads to the formation of ATP from adenosine diphosphate (ADP)

ATP diffuses throughout the cell to provide energy for other cellular work, such as biosynthetic reactions, ion transport, and cell movement. The chemi¬ cal potential energy of ATP is made available when it transfers one (or two) of its phosphate groups to another molecule. This process can be represented by the reverse of the preceding reaction, namely, the hydrolysis of ATP to ADP. (As we will see in later chapters, the interconversion of ATP and ADP in the cell is not freely reversible, and free phosphate groups are seldom released di¬ rectly from ATP.) The degree to which ATP participates in routine cellular ac¬ tivities is illustrated by calculations indicating that while the concentration of cellular ATP is relatively moderate (~5 mM), humans typically recycle their own weight of ATP each day. Nucleotide derivatives participate in a wide variety of metabolic processes. For example, starch synthesis in plants proceeds by repeated additions of glu¬ cose units donated by ADP-glucose (Fig. 3-2). Other nucleotide derivatives, as we will see in later chapters, carry groups that undergo oxidation-reduction reactions. The attached group, which may be a small molecule such as glucose (Fig. 3-2) or even another nucleotide, is typically linked to the

nucleotide through a mono- or diphosphate group. Nucleotides can be joined to each other to form the polymers that are familiar to us as RNA and DNA. In this section, we describe the general features of these nucleic acids. Nucleic acid structure is considered further in Chapter 24.

A Nucleic Acids Are Polymers of Nucleotides

Nucleic acids are chains of nucleotides whose phosphates bridge the 3' and 5' positions of neighboring ribose units (Fig. 3-3). The phosphates of these polynucleotides are acidic, so at physiological pH, nucleic acids are polyanions. The linkage between individual nucleotides is known as a phos¬ phodiester bond, so named because the phosphate is esterified to two ribose units. Each nucleotide that has been incorporated into the polynucleotide is known as a nucleotide residue. The terminal residue whose C5' is not linked to another nucleotide is called the 5 ' end, and the terminal residue whose C3' is not linked to another nucleotide is called the 3' end. By convention, the sequence of nucleotide residues in a nucleic acid is written, left to right, from the 5' end to the 3' end.

The properties of a polymer such as a nucleic acid may be very different from the properties of the individual units, or monomers, before polymeriza¬ tion. As the size of the polymer increases from dimer, trimer, tetramer, and so on through oligomer (Greek: oligo, few), physical properties such as charge and solubility may change. In addition, a polymer of nonidentical residues has a property that its component monomers lack — namely ; it contains information in the form of its sequence of residues.

Chargaff’s Rules Describe the Base Composition of DNA. Although there appear to be no rules governing the nucleotide composition of typical RNA molecules, DNA has equal numbers of adenine and thymine residues (A = T) and equal numbers of guanine and cytosine

residues (G = C). These relation¬ ships, known as ChargafFs rules, were discovered in the late 1940s by Erwin Chargaff, who devised the first reliable quantitative methods for the composi¬ tional analysis of DNA. DNA base composition varies widely among different organisms. It ranges from ~25 to 75 mol % G + C in different species of bacteria. However, it is more or less constant among related species; for example, in mammals G + C ranges from 39 to 46%. The significance of Chargaff s rules was not immediately appreciated, but we now know that the structural basis for the rules derives from DNAs double-stranded nature.

DNA Forms a Double Helix

The determination of the structure of DNA by James Watson and Francis Crick in 1953 is often said to mark the birth of modern molecular biology. The Watson-Crick structure of DNA not only provided a model of what is arguably the central molecule of life, it also suggested the molecular mecha¬ nism of heredity. Watson and Cricks accomplishment, which is ranked as one of sciences major intellectual achievements, was based in part on two pieces of evidence in addition to Chargaff s rules: the correct tautomeric forms of the bases and indications that DNA is a helical molecule.

The purine and pyrimidine bases of nucleic acids can assume different tautomeric forms. tautomers are easily converted isomers that differ only in hydrogen positions. X-Ray, nuclear magnetic resonance (NMR), and spectroscopic investigations have firmly established that the nucleic acid bases are overwhelmingly in the keto tautomeric forms . In 1953, however, this was not generally appreciated. Information about the dominant tautomeric forms was provided by Jerry Donohue, an office mate of Watson and Crick and an expert on the X-ray structures of small organic molecules.

Evidence that DNA is a helical molecule was provided by an X-ray dif¬ fraction photograph of a DNA fiber taken by Rosalind Franklin. The appearance of the photograph enabled Crick, an X-ray crystallographer

by training, to deduce (a) that DNA is a helical molecule and (b) that its planar aromatic bases form a stack that is parallel to the fiber axis.

The limited structural information, along with Chargaff's rules, provided few clues to the structure of DNA; Watson and Cricks model sprang mostly from their imaginations and model-building studies. Once the Watson-Crick model had been published, however, its basic simplicity combined with its ob¬ vious biological relevance led to its rapid acceptance. Later investigations have confirmed the general validity of the Watson-Crick model, although its de¬ tails have been modified.

The Watson-Crick model of DNA has the following major features:

1. Two polynucleotide chains wind around a common axis to form a double helix.

2. The two strands of DNA are antiparallel (run in opposite directions), but each forms a right-handed helix.

3. The bases occupy the core of the helix and sugar—phosphate chains run along the periphery, thereby minimizing the repulsions between charged phosphate groups. The surface of the double helix contains two grooves of unequal width: the major and minor grooves.

4. Each base is hydrogen bonded to a base in the opposite strand to form a planar base pair. The Watson—Crick structure can accommodate only two types of base pairs. Each adenine residue must pair with a thymine residue and vice versa, and each guanine residue must pair with a cytosine residue and vice versa (Fig. 3-8). These hydrogen-bonding interactions, a phenomenon known as complementary base pairing, result in the specific association of the two chains of the double helix.

The Watson—Crick structure can accommodate any sequence of bases on one polynucleotide strand if the opposite strand has the complementary base sequence. This immediately accounts for Chargaff s

rules. More importantly, it suggests that each DNA strand can act as a template for the synthesis of its complementary strand and hence that hereditary information is encoded in the sequence of bases on either strand.

Most DNA Molecules Are Large. The extremely large size of DNA molecules is in keeping with their role as the repository of a cells genetic information. Of course, an organisms genome, its unique DNA content, may be allocated among several chromosomes (Greek: chromos , color + soma, body), each of which contains a separate DNA molecule. Note that many organisms are diploid; that is, they contain two equivalent sets of chromosomes, one from each parent. Their content of unique (haploid) DNA is half their total DNA. For example, humans are diploid organisms that carry 46 chromosomes per cell; their haploid number is therefore 23.

Because of their great lengths, DNA molecules are described in terms of the number of base pairs (bp) or thousands of base pairs (kilobase pairs, or kb). Naturally occurring DNAs vary in length from ~5 kb in small DNAcontaining viruses to well over 250,000 kb in the largest mammalian chromo¬ somes. Although DNA molecules are long and relatively stiff, they are not completely rigid. We will see in Chapter 24 that the DNA double helix forms.

coils and loops when it is packaged inside the cell. Furthermore, depending on the nucleotide sequence, DNA may adopt slightly different helical conformations. Finally, in the presence of other cellular components, the DNA may bend sharply or the two strands may partially unwind.

RNA Is a Single-Stranded Nucleic Acid

Single-stranded DNA is rare, occurring mainly as the hereditary material of certain viruses. In contrast, RNA occurs primarily as single strands, which usu¬ ally form compact structures rather than loose extended

chains (double stranded RNA is the hereditary material of certain viruses). An RNA strand—which is identical to a DNA strand except for the presence of 2'-OH groups and the substitution of uracil for thymine—can base-pair with a complementary strand of RNA or DNA. As expected, A pairs with U (or T in DNA), and G with C. Base pairing often occurs intramolecularly, giving rise to stem-loop structures or, when loops interact with each other, to more complex structures.

The intricate structures that can potentially be adopted by single-stranded RNA molecules provide additional evidence that RNA can do more than just store and transmit genetic information. Numerous investigations have found that certain RNA molecules can specifically bind small organic molecules and can catalyze reactions involving those molecules. These findings provide sub¬ stantial support for theories that many of the processes essential for life began through the chemical versatility of small polynucleotides (a situation known as the RNA world).

Overview of Nucleic Acid Function

KEY CONCEPTS

• DNA carries genetic information in the form of its sequence of nucleotides.

• The nucleotide sequence of DNA is transcribed into the nucleotide sequence of messenger RNA, which is then translated into a protein, a sequence of amino acids.

DNA is the carrier of genetic information in all cells and in many viruses. Yet a period of over 75 years passed from the time the laws of inheritance were discovered by Gregor Mendel until the biological role of DNA was elucidated. Even now, many details of how genetic information is expressed and transmit¬ ted to future generations remain unclear.

Mendels work with garden peas led him to postulate that an individual plant contains a pair of factors (which we now call genes), one inherited from each parent. But Mendels theory of inheritance, reported in 1866, was almost universally ignored by his contemporaries, whose knowledge of anatomy and physiology provided no basis for its understanding. Eventually, genes were hy¬ pothesized to be part of chromosomes, and the pace of genetic research greatly accelerated.

A DNA Carries Genetic Information

Until the 1940s, it was generally assumed that genes were made of protein, since proteins were the only biochemical entities that, at the time, seemed complex enough to serve as agents of inheritance. Nucleic acids, which had first been isolated in 1869 by Friedrich Miescher, were believed to have mo¬ notonously repeating nucleotide sequences and were therefore unlikely candi¬ dates for transmitting genetic information.

It took the efforts of Oswald Avery, Colin MacLeod, and Maclyn McCarty to demonstrate that DNA carries genetic information. Their experiments, com¬ pleted in 1944, showed that DNA—not protein—extracted from a virulent (pathogenic) strain of the bacterium Diplococcus pneumoniae was the substance that transformed (permanently changed) a nonpathogenic strain of the organ¬ ism to the virulent strain (Fig. 3-10). Avery's discovery was initially greeted with skepticism, but it influenced Erwin Chargaff, whose rules (Section 3-2A) led to subsequent models of the structure and function of DNA.

The double-stranded, or duplex, nature of DNA facilitates its replication. When a cell divides, each DNA strand acts as a template for the assembly of its complementary strand (Fig. 3-11). Consequently, every progeny cell con¬ tains a complete DNA molecule (or a complete set of DNA molecules in or¬ ganisms whose genomes contain more than one chromosome). Each DNA molecule consists of one parental strand and one daughter strand. Daughter strands are synthesized by the stepwise polymerization of nucleotides that specifically pair with bases on the parental strands. The mechanism of repli¬ cation, while straightforward in principle, is exceedingly complex in the cell, requiring a multitude of cellular factors to proceed with fidelity and efficiency, as we will see in Chapter 25.

B Genes Direct Protein Synthesis

The question of how sequences of nucleotides control the characteristics of or¬ ganisms took some time to be answered. In experiments with the mold Neurospora crassa in the 1940s, George Beadle and Edward Tatum found that there is a specific connection between genes and enzymes , the one gene—one enzyme theory Beadle and Tatum showed that mutant varieties of Neurospora that were generated by irradiation with X-rays

required additional nutrients in order to grow. Presumably, the offspring of the radiation-damaged cells lacked the spe¬ cific enzymes necessary to synthesize those nutrients.

The link between DNA and enzymes (nearly all of which are proteins) is RNA. The DNA of a gene is transcribed to produce an RNA molecule that is complementary to the DNA. The RNA sequence is then translated into the cor¬ responding sequence of amino acids to form a protein (Fig. 3-12). These trans¬ fers of biological information are summarized in the so-called central dogma of molecular biology formulated by Crick in 1958.

In this diagram, arrows represent information transfer when DNA directs its own replication to produce new DNA molecules; when DNA is transcribed into RNA; and when RNA is translated into protein.

Just as the daughter strands of DNA are synthesized from free deoxynucleoside triphosphates that pair with bases in the parent DNA strand, RNA strands are synthesized from free ribonucleoside triphosphates that pair with the complementary bases in one DNA strand of a gene (transcription is de¬ scribed in greater detail in Chapter 26). The RNA that corresponds to a protein-coding gene (called messenger RNA, or mRNA) makes its way to a ribosome, an organelle that is itself composed largely of RNA (ribosomal RNA, or rRNA). At the ribosome, each set of three nucleotides in the mRNA pairs with three complementary nucleotides in a small RNA molecule called a transfer RNA, or tRNA (Fig. 3-13). Attached to each tRNA molecule is its corresponding amino acid. The ribosome catalyzes the joining of amino acids, which are the monomeric units of proteins (protein synthesis is described in detail in Chapter 27). Amino acids are added to the growing protein chain according to the order in which the tRNA molecules bind

to the mRNA. Since the nucleotide sequence of the mRNA in turn reflects the sequences of nu¬ cleotides in the gene, DNA directs the synthesis of proteins. It follows that al¬ terations to the genetic material of an organism (mutations) may manifest themselves as proteins with altered structures and functions.

Using techniques that are described in the following sections and in other parts of this book, researchers can compile a catalog of all the information en¬ coded in an organisms DNA. The study of the genomes size, organization, and gene content is known as genomics. By analogy, transcriptomics refers to the study of gene expression, which focuses on the set of mRNA molecules, or transcriptome, that is transcribed from DNA under any particular set of circumstances. Finally, proteomics is the study of the proteins (the proteome) produced as a result of transcription and translation. Although an organisms genome remains essentially unchanged throughout its lifetime, its transcrip¬ tome and proteome may vary significantly among different types of tissues, developmental stages, and environmental conditions.

4 Nucleic Acid Sequencing

KEY

CONCEPTS

• In the laboratory, nucleic acids can be cut at specific sequences by restriction enzymes.

• Nucleic acid fragments are separated by size using electrophoresis.

• In the chain-termination method, DNA polymerase generates DNA fragments that are randomly terminated. The identities of the terminator nucleotides of successive fragments reveal the original DNA sequence.

• The human genome contains —23,000 genes, corresponding to about 1.2% of its 3 billion nucleotides.

• Sequence differences reveal evolutionary changes.

Much of our current understanding of protein structure and function rests squarely on information gleaned not from the proteins themselves, but indi¬ rectly from their genes. The ability to determine the sequence of nucleotides innucleic acids has made it possible to deduce the amino acid sequences of their en¬ coded proteins and, to some extent, the structures and functions of those proteins. Nucleic acid sequencing has also revealed information about the regulation of genes. Certain portions of genes that are not actually transcribed into RNA neverthe¬ less influence how often a gene is transcribed and translated, that is, expressed. Moreover, efforts to elucidate the sequences in hitherto unmapped regions of DNA have led to the discovery of new genes and new regulatory elements. Once in hand, a nucleic acid sequence can be duplicated, modified, and expressed, making it possible to study proteins that could not otherwise be obtained in useful quantities. In this section, we describe how nucleic acids are sequenced and what information the sequences may reveal. In the following section, we dis¬ cuss the manipulation of purified nucleic acid sequences for various purposes.

The overall strategy for sequencing any polymer of nonidentical units is

1. Cleave the polymer into fragments that are small enough to be fully sequenced.

2. Determine the sequence of residues in each fragment.

3. Determine the order of the fragments in the original polymer by align¬ ing fragments that contain overlapping sequences.

The first efforts to sequence RNA used nonspecific enzymes to generate rela¬ tively small fragments whose nucleotide composition was then determined by partial digestion with an enzyme that selectively removed nucleotides from one end or the other (Fig. 3-14). Sequencing RNA in this manner was tedious and time-consuming. Using such methods, it took Robert Holley 7 years to determine the sequence of a 76-residue tRNA molecule.

After 1975, dramatic progress was made in nucleic acid sequencing tech¬ nology. The advances were made possible by the discovery of enzymes that could cleave DNA at specific sites and by the development of rapid sequenc¬ ing techniques for DNA. Because most specific DNA sequences are normally present in a genome in only a single copy, most sequencing projects take ad¬ vantage of methods to amplify segments of DNA by cloning or copying them (Section 3-5).

A Restriction Endonucleases Cleave DNA at Specific Sequences

Many bacteria are able to resist infection by bacteriophages (viruses that are specific for bacteria) by virtue of a restriction-modification system. The bac¬ terium modifies certain nucleotides in specific sequences of its own DNA by adding a methyl (—CH 3) group in a reaction catalyzed by a modification methylase. A restriction endonuclease, which recognizes the same nucleotide sequence as does the methylase, cleaves any DNA that has not been modified on at least one of its two strands. (An endonuclease cleaves a nucleic acid within the polynucleotide strand; an exonuclease cleaves a nucleic acid by re¬ moving one of its terminal residues.) This system destroys foreign (phage) DNA containing a recognition site that has not been modified by methylation. The host DNA is always at least half methylated, because although the daughter strand is not methylated until shortly after it is synthesized, the parental strand to which it is paired is already modified (and thus protects both strands of the DNA from cleavage by the restriction enzyme).

Type II restriction endonucleases are particularly useful in the laboratory. These enzymes cleave DNA within the four- to eight-base sequence that is recognized by their corresponding modification methylase. (Type I and Type III restriction endonucleases cleave DNA at sites other than their recognition sequences.) Nearly 4000 Type II restriction enzymes with over 270 different recognition sequences have been characterized. Some of the more widely used restriction enzymes are listed in Table 3-2. A restriction enzyme is named by the first letter of the genus and the first two letters of the species of the bac¬ terium that produced it, followed by its serotype or strain designation, if any, and a roman numeral if the bacterium contains more than one type of restric¬ tion enzyme. For example, EcoRI is produced by E. coli strain RY13.

Interestingly, most Type II restriction endonucleases recognize and cleave palindromic DNA sequences. A palindrome is a word or phrase that reads the same forward or backward. Two examples are “refer” and “Madam, I’m Adam.” In a palindromic DNA segment, the sequence of nucleotides is the same in each strand, and the segment is said to have twofold symmetry (Fig. 3-15). Most restriction enzymes cleave the two strands of DNA at positions that are staggered, producing DNA fragments with complementary single-strand exten¬ sions. Restriction fragments with such sticky ends can associate by base pairing with other restriction fragments generated by the same restrictionenzyme. Some restriction endonucleases cleave the two strands of DNA at the symmetry axis to yield restriction fragments with fully base-paired blunt ends.

B Electrophoresis Separates Nucleic Acids According to Size

Treating a DNA molecule with a restriction endonuclease produces a series of precisely defined fragments that can be separated according to size. Gel elec¬ trophoresis is commonly used for the separation. In principle, a charged mol¬ ecule moves in an electric field with a velocity proportional to its overall charge density, size, and shape. For molecules with a relatively homogeneous compo¬ sition (such as nucleic acids), shape and charge density are constant, so the velocity depends primarily on size. Electrophoresis is carried out in a gel-like matrix, usually made from agarose (carbohydrate polymers that form a loose mesh) or polyacrylamide (a more rigid cross-linked synthetic polymer). The gel is typically held between two glass plates (Fig. 3-16) or inside a narrow capillary tube. The molecules to be separated are applied to one end of the gel, and the molecules move through the pores in the matrix under the influ¬ ence of an electric field. Smaller molecules move more rapidly through the gel and therefore migrate farther in a given time.

Following electrophoresis, the separated molecules may be visualized in the gel by an appropriate technique, such as addition of a stain that binds tightly to the DNA, by radioactive labeling, or by their fluorescence. Depending on the dimensions of the gel and the visualization technique used, samples containing less than a nanogram of material can be separated and de¬ tected by gel electrophoresis. Several samples can be electrophoresed simulta¬ neously. For example, the fragments obtained by digesting a DNA sample with different restriction endonucleases can be visualized side by side (Fig. 3-17).

The sizes of the various fragments can be determined by comparing their electrophoretic mobilities to the mobilities of fragments of known size.

C Traditional DNA Sequencing Uses the Chain-Terminator Method

Until recent years, the most widely used technique for sequencing DNA was the chain-terminator method, which was devised by Frederick Sanger. The first step in this procedure is to obtain single polynucleotide strands. Complementary DNA strands can be separated by heating, which breaks the hydrogen bonds between bases. Next, polynucleotide fragments that terminate at positions corresponding to each of the four nucleotides are generated. Finally, the fragments are separated and detected.

DNA Polymerase Copies a Template Strand. The chain-terminator method (also called the dideoxy method) uses an E. coli enzyme to make complementary copies of the single-stranded DNA being sequenced. The en¬ zyme is a fragment of DNA polymerase I, one of the enzymes that partic¬ ipates in replication of bacterial DNA (Section 25-2A). Using the

single DNA strand as a template, DNA polymerase I assembles the four deoxynucleoside triphosphates (dNTPs), dATP, dCTP, dGTP, and dTTP, into a com¬ plementary polynucleotide chain that it elongates in the 5' —> 3' direction (Fig. 3-18).

DNA polymerase I can sequentially add deoxynucleotides only to the 3' end of a polynucleotide. Hence, replication is initiated in the presence of a short polynucleotide (a primer) that is complementary to the 3' end of the template DNA and thus becomes the 5' end of the new strand. The primer base-pairs with the template strand, and nucleotides are sequentially added to the 3' end of the primer. If the DNA being sequenced is a restriction frag¬ ment, as it usually is, it begins and ends with a restriction site. The primer can therefore be a short DNA segment with the sequence of this restriction site.

DNA Synthesis Terminates after Specific Bases. In the chain-terminator technique (Fig. 3-19), the DNA to be sequenced is incubated with DNA poly¬ merase I, a suitable primer, and the four dNTP substrates (reactants in enzy¬ matic reactions) for the polymerization reaction. The key component of the reaction mixture is a small amount of a 2',3'-dideoxynucleoside triphos¬ phate (ddNTP),

which lacks the 3'-OH group of deoxynucleotides. When the dideoxy analog is incorporated into the growing polynucleotide in place of the corresponding normal nucleotide, chain growth is terminated because addition of the next nucleotide re¬ quires a free 3'-OH.

By using only a small amount of the ddNTP, a series of truncated chains is generated, each of which ends with the dideoxy analog at one of the positions occupied by the corresponding base. Each ddNTP bears a different fluorescent “tag” so that the products of the polymerase reaction can be readily detected. Gel electrophoresis separates the newly

synthesized DNA segments, which differ in size by one nucleotide. Thus, the sequence of the replicated strand can be di¬ rectly read from the gel. Note that the sequence obtained by the chain-termi¬ nator method is complementary to the DNA strand being sequenced.

The most advanced sequencing devices that employ the chain-terminator method identify each DNA fragment as it exits the bottom of a capillary electrophoresis tube, so the sequence data take the form of a series of peaks (Fig. 3-20). Sample preparation and data analysis are fully automated, and sequences of DNA up to 1000 nucleotides can be obtained from a single re¬ action mixture. Moreover, these systems contain arrays of 96 capillary tubes and hence can simultaneously sequence 96 DNA segments.

Newer Sequencing Technologies Use Light or Voltage Changes. The value of DNA sequence information, particularly the enormous datasets obtained by sequencing entire genomes, has driven the development of novel sequencing technologies that offer various trade-offs among cost, speed, and accuracy. Like the older chain-terminator procedure, newer methods take advantage of the ability of DNA polymers to produce a complementary copy of a template DNA strand.

In pyrosequencing, molecules of the template DNA are immobilized on the surfaces of microscopic plastic beads that are deposited in small wells in a fiber-optic slide with one bead per well. A primer and DNA polymerase are added, and then a dNTP substrate is introduced. If DNA polymerase adds that nucleotide to the new DNA strand, pyrophosphate is released and trig¬ gers a chemical reaction involving the firefly enzyme luciferase, which gener¬ ates a flash of light. Solutions of each of the four dNTPs are successively washed across the immobilized DNA

template, and a detector records whether light is produced in the presence of a particular dNTP. In this way, the se¬ quence of nucleotides complementary to the template strand can be deduced, and no electrophoretic separation is needed. Pyrosequencing can accurately "read" stretches of 300—500 nucleotides, somewhat shorter than the sequences revealed by dideoxy sequencing. The fiber-optic slides contain numerous wells so that as many as ~400,000 templates can be sequenced simultaneously. Consequently, the pyrosequencing system is ~300-fold faster than the most advanced dideoxy sequencing systems.

Other sequencing instruments detect the proton generated on pyrophos¬ phate release, so that a DNA sequence can be deduced from the minute change in voltage when the correct nucleotide (the one matching the template at that point) is incorporated into the new chain. All of these devices, like the pyrose¬ quencing system, sequence large numbers of DNA fragments simultaneously and hence determine the identities of billions of nucleotides in a single run.

Databases Store Nucleotide Sequences. The results of sequencing projects large and small are customarily deposited in online databases such as GenBank (see Bioinformatics Project 1). Approximately 300 billion nucleotides represent¬ ing over 200 million sequences have been recorded as of late 2010.

Box 3-1 Pathways of Discovery

Francis Collins and the Gene for Cystic Fibrosis

Francis S. Collins (1950-) By the mid¬ twentieth century, the molecular bases of several human diseases were appreciated. For example, sickle-cell anemia (Section 7-1 E) was known to be caused by an abnormal hemoglobin protein. Studies of sickle-cell hemoglobin eventually revealed the underlying genetic defect, a mutation in a hemoglobin gene. It therefore seemed possible to trace other diseases to defective genes. But for many genetic diseases, even those with well-characterized symptoms, no defective protein had yet been identified. One such disease was cystic fibrosis, which is characterized mainly by the secretion of thick mucus that obstructs the airways and creates an ideal environment for bacterial growth. Cystic fibrosis is the most common inherited disease in individuals of northern European descent, striking about 1 in 2500 newborns and leading to death by early adulthood due to irreversible lung damage. It was believed that identifying the molecular defect in cystic fibrosis would lead to better understanding of the disease and to the ability to design more effective treatments.

Enter Francis Collins, who began his career by earning a doctorate in physical chemistry but then enrolled in medical school to take part in the molecular biology revolution. As a physicianscientist, Collins developed methods for analyzing large stretches of DNA in order to home in on specific genes, including the one that, when mutated, causes cystic fibrosis. By analyzing the DNA of individuals with the disease (who had two copies of the defective gene) and of family members who were asymptomatic carriers (with one normal and one defective copy of the gene), Collins and his team localized the cystic fibrosis gene to the long

arm of chromosome 7. They gradually closed in on a DNA segment that appears to be present in a number of mammalian species, which suggests that the segment contains an essential gene. The cystic fibrosis gene was finally identified in 1989. Collins had demonstrated the feasibility of identifying a genetic defect in the absence of other molecular information.

Once the cystic fibrosis gene was in hand, it was a relatively straightforward process to deduce the probable structure and function of the encoded protein, which turned out to be a membrane channel for chloride ions. When functioning normally, the protein helps regulate the ionic composition and viscosity of extracellular secretions. Discovery of the cystic fibrosis gene also made it possible to design tests to identify carriers so that they could take advantage of genetic counseling.

Throughout Collins' work on the cystic fibrosis gene and during subsequent hunts for the genes that cause neurofibromatosis and Huntington's disease, he was mindful of the ethical implications of the new science of molecular genetics. Collins has been a strong advocate for protecting the privacy of genetic information. At the same time, he recognizes the potential therapeutic use of such information. In his tenure as director of the human genome project, he was committed to making the results freely and immediately accessible, as a service to researchers and the individuals who might benefit from new therapies based on molecular genetics. He is presently the director of the National Institutes of Health (NIH).

Riordan, J.R., Rommens, J.M., Kerem, B.-S., Alon, N., Rozmahel, R., Grzelczak, Z., Zielensky, J., Lok, S., Plavsic, N., Chou, J.-L, Drumm, M.L, lannuzzi, M.C., Collins, F.S., and Tsui, L-C., Identification of the

cystic fibrosis gene: Cloning and characterization of complementary DNA, Science 245 , 1066-1073 (1989).

Nucleic acid sequencing has become so routine that directly determining a proteins amino acid sequence (Section 5-3) is generally far more timeconsuming than determining the base sequence of its corresponding gene. In fact, nucleic acid sequencing is invaluable for studying genes whose products have not yet been identified. If the gene can be sequenced, the probable func¬ tion of its protein product may be deduced by comparing the base sequence to those of genes whose products are already characterized (see Box 3-1).

D Entire Genomes Have Been Sequenced

The advent of large-scale sequencing techniques brought to fruition the dream of sequencing entire genomes. However, the major technical hurdle in se¬ quencing all the DNA in an organisms genome is not the DNA sequencing itself but, rather, assembling the tens of thousands to tens of millions of se¬ quenced segments (depending on the size of the genome) into contiguous blocks and assigning them to their correct chromosomal positions. To do so required the development of automated sequencing protocols and mathemat¬ ically sophisticated computer algorithms.

The first complete genome sequence to be determined, that of the bacterium Haemophilus influenzae, was reported in 1995 by Craig Venter. By mid-2011, the complete genome sequences of over 1600 prokaryotes had been reported (with many more being determined) as

well as over 140 eukaryotes, including humans, human pathogens, plants, and laboratory organisms (Table 3-3).

In metagenomic sequencing, the DNA sequences of multiple organisms are analyzed as a single data set. This approach is used to characterize com¬ plex microbial communities, such as those in marine environments, where in¬ dividual species—including many not yet identified—cannot be cultured and sequenced one by one. Metagenomic sequence data reveal the overall gene number and an estimate of the collective metabolic capabilities of the com¬ munity. Over 3 million genes have been identified in a metagenomic analysis of the microorganisms that inhabit the human gut, representing some 1000 bacterial species. While most humans share a common core set of about 60 gut microorganisms, significant differences appear to correlate with metabolic variables such as body mass.

The Human Genome Contains Relatively Few Genes. The determination of the 3-billion-nucleotide human genome sequence was a gargantuan under¬ taking involving hundreds of scientists working in two groups, one led by Venter and the other by Francis Collins (Box 3-1), Eric Lander, and John Sulston. After over a decade of intense effort, the "rough draft" of the human genome sequence was reported in early 2001 and the "finished" sequence, cov¬ ering —99% of the genome, was reported in 2004. This stunning achievement promises to revolutionize the way both biochemistry and medicine are viewed and practiced, although it is likely to require many years of further effort before its full significance is understood. Nevertheless, numerous important conclusions can already be drawn, including these:

1. About half the human genome consists of repeating sequences of various types.

2. Up to 80% of the genome may be transcribed to RNA.

3. Only 1.2% of the genome encodes protein.

4. The human genome appears to contain only —23,000 protein-encoding genes [also known as open reading frames (ORFs)] rather than the 50,000 to 140,000 ORFs that had previously been predicted. This com¬ pares with the —6000 ORFs in yeast, —13,000 in Drosophila. , —19,000 in C. elegans, and —26,000 in Arabidopsis (although these numbers will almost certainly change as our ability to recognize ORFs improves).

5. Only a small fraction of human proteins are unique to vertebrates; most occur in other if not all life-forms.

6. Two randomly selected human genomes differ, on average, by only 1 nucleotide per 1000; that is, any two people are likely to be >99.9% genetically identical. The obviously greater complexity of humans (vertebrates) relative to inverte¬ brate forms of life is unlikely to be due to the not-much-larger numbers of ORFs that vertebrates encode. Rather, it appears that vertebrate proteins them¬ selves are more complex than those of invertebrates; that is, vertebrate proteins tend to have more domains (modules) than invertebrate proteins, and these modules are more often selectively expressed through alternative gene splic¬ ing (a phenomenon in which a given gene transcript can be processed in multiple ways so as to yield different proteins when translated; Section 26-3B). In fact, most vertebrate genes encode several different although similar proteins.

E Evolution Results from Sequence Mutations

One of the richest rewards of nucleic acid sequencing technology is the infor¬ mation it provides about the mechanisms of evolution. The chemical and physical properties of DNA, such as its regular three-dimensional shape and the elegant process of replication, may leave the impression that genetic infor¬ mation is relatively static. In fact, DNA is a dynamic molecule , subject to changes that alter genetic information. For example, the mispairing of bases during DNA replication can introduce errors known as point mutations in the daughter strand. Mutations also result from DNA damage by chemicals or ra¬ diation. More extensive alterations in genetic information are caused by faulty recombination (exchange of DNA between chromosomes) and the transpo¬ sition of genes within or between chromosomes and, in some cases, from one organism to another. All these alterations to DNA provide the raw material for natural selection. When a mutated gene is transcribed and the messenger RNA is subsequently translated, the resulting protein may have properties that confer some advantage to the individual. As a beneficial change is passed from generation to generation, it may become part of the standard genetic makeup of the species. Of course, many changes occur as a species evolves, not all of them simple and not all of them gradual.

Phylogenetic relationships can be revealed by comparing the sequences of similar genes in different organisms. The number of nucleotide differences be¬ tween the corresponding genes in two species roughly indicates the degree to which the species have diverged through evolution. The regrouping of prokary¬ otes into archaea and bacteria (Section 1-2C) according to rRNA sequences present in all organisms illustrates the impact of sequence analysis.

Nucleic acid sequencing also reveals that species differing in phenotype (physical characteristics) are nonetheless remarkably similar at the molecular level. For example, humans and chimpanzees share nearly 99% of their DNA.

Studies of corn (maize) and its putative ancestor, teosinte, suggest that the plants differ in only a handful of genes governing kernel development (teosinte kernels are encased by an inedible shell; Fig. 3-21).

Small mutations in DNA are apparently responsible for relatively large evo¬ lutionary leaps. This is perhaps not so surprising when the nature of genetic in¬ formation is considered. A mutation in a gene segment that does not encode protein might interfere with the binding of cellular factors that influence the timing of transcription. A mutation in a gene encoding an RNA might inter¬ fere with the binding of factors that affect the efficiency of translation. Even a minor rearrangement of genes could disrupt an entire developmental process, resulting in the appearance of a novel species. Notwithstanding the high prob¬ ability that most sudden changes would lead to diminished individual fitness or the inability to reproduce, the capacity for sudden changes in genetic informa¬ tion is consistent with the fossil record. Ironically, the discontinuities in the fos¬ sil record that are probably caused in part by sudden genetic changes once fueled the adversaries of Charles Darwins theory of evolution by natural selection.

T Sequence Variations Can Be Linked to Human Diseases. Almost 2000 genes have been linked to different human diseases, yet monogenetic dis¬ eases, such as cystic fibrosis (see Box 3-1), are relatively rare. Most diseases re¬ sult from interactions among multiple genes and from

environmental factors. Nevertheless, scientists are hopeful that genomics can lead to a better under¬ standing of how genetic information impacts human health and susceptibility to illness. One area of success is the screening of individuals who are carriers for a recessive genetic disorder; that is, they have a normal phenotype but bear one copy of the defective gene, which may be passed to their children. Clinical tests are available to detect over a hundred such single-gene defects (Table 3-4). Preliminary results suggest that as more parents become aware of their carrier status, fewer children with the disease are being born.

Although new sequencing technologies will soon make it feasible to se¬ quence the complete genomes of individuals, for example, members of a fam¬ ily with a poorly characterized genetic disease, most approaches focus on single-nucleotide polymorphisms (SNPs, instances where the DNA sequence differs among individuals at one nucleotide). From studies involving thousands of subjects with and without certain complex diseases such as cancer and type 2 diabetes, researchers have identified a number of SNPs that are associated with increased risk for these conditions. The SNPs are only proxies for genes but could provide starting points for further efforts to locate genes or regula¬ tory DNA sequences that might be directly involved in the disease. An ongo¬ ing challenge for this work is that each genetic variant associated with a disease typically increases risk for the disease by only a few percent, so an individuals likelihood of developing a particular disease appears to be a complicated func¬ tion of which variants are present. Several commercial enterprises offer indi¬ vidual genome-sequencing services, but until genetic information can be reliably translated into effective disease-prevention or treatment regimens, the practical value of “personal genomics” is quite limited.

5 Manipulating DNA

KEY

CONCEPTS

• Segments of DNA can be cloned, or reproduced, in a host organism.

• A DNA library is a collection of cloned DNA segments that can be screened to find a particular gene.

• The polymerase chain reaction amplifies a DNA segment by repeatedly synthesizing complementary strands.

• Recombinant DNA technology can be used to manipulate genes for protein expression or for the production of transgenic organisms.

Along with nucleic acid sequencing, techniques for manipulating DNA in vitro and in vivo (in the test tube and in living systems) have produced dra¬ matic advances in biochemistry, cell biology, and genetics. In many cases, this recombinant DNA technology has made it possible to purify specific DNA sequences and to prepare them in quantities sufficient for study. Consider the problem of isolating a unique 1000-bp length of chromosomal DNA from E. coli. A 10-L culture of cells grown at a density of —10 10 cells • mL 1 con¬ tains only —0.1 mg of the desired DNA, which would be all but impossible to separate from the rest of the DNA using classical separation techniques (Sections 5-2 and 24-3). Recombinant DNA technology, also called molecular cloning or genetic engineering makes it possible to isolate, amplify, and mod¬ ify specific DNA sequences.

A Cloned DNA Is an Amplified Copy

The following approach is used to obtain and amplify a segment of DNA:

1. A fragment of DNA of the appropriate size is generated by a restric¬ tion enzyme, by PCR (Section 3-5C), or by chemical synthesis.

2. The fragment is incorporated into another DNA molecule known as a vector, which contains the sequences necessary to direct DNA replication.

3. The vector—with the DNA of interest—is introduced into cells, in which it is replicated.

4. Cells containing the desired DNA are identified, or selected.

Cloning refers to the production of multiple identical organisms derived from a single ancestor. The term clone refers to the collection of cells that contain the vector carrying the DNA of interest or to the DNA itself.

In a suitable host organism, such as E. coli or yeast, large amounts of the inserted DNA can be produced.

Cloned DNA can be purified and sequenced (Section 34). Alternatively, if a cloned gene is flanked by the properly positioned regulatory sequences for RNA and protein synthe¬ sis, the host may also produce large quantities of the RNA and protein specified by that gene. Thus, cloning provides materials (nucleic acids and proteins) for other studies and also provides means for studying gene expression under controlled conditions.

Cloning Vectors Carry Foreign DNA. A variety of small, autonomously repli¬ cating DNA molecules are used as cloning vectors.

Plasmids are circular DNA molecules of 1 to 200 kb found in bacteria or yeast cells. Plasmids can be con¬ sidered molecular parasites, but in many instances they benefit their host by providing functions, such as resistance to antibiotics, that the host lacks.

Some types of plasmids are present in one or a few copies per cell and replicate only when the bacterial chromosome replicates. However, the plas¬ mids used for cloning are typically present in hundreds of copies per cell and can be induced to replicate until the cell contains two or three thousand copies (representing about half of the cells total DNA). The plasmids that have been constructed for laboratory use are relatively small, replicate easily, carry genes specifying resistance to one or more antibiotics, and contain a number of con¬ veniently located restriction endonuclease sites into which foreign DNA can be inserted. Plasmid vectors can be used to clone DNA segments of no more than ~10 kb. The E. coli plasmid designated pUC18 (Fig. 3-22) is a repre¬ sentative cloning vector (“pUC” stands for plasmid-Universal Cloning).

Bacteriophage \ (Fig. 3-23) is an alternative cloning vector that can ac¬ commodate DNA inserts up to 16 kb. The central third of the 48.5-kb phage genome is not required for infection and can therefore be replaced

by foreign DNAs of similar size. The resulting recombinant, or chimera (named after the mythological monster with a lions head, goats body, and serpents tail), is packaged into phage particles that can then be introduced into the host cells. One advantage of using phage vectors is that the recombinant DNA is pro¬ duced in large amounts in easily purified form. Baculoviruses, which infect insect cells, are similarly used for cloning in cultures of insect cells.

Much larger DNA segments—up to several hundred kilobase pairs—can be cloned in large vectors known as bacterial artificial chromosomes (BACs) or yeast artificial chromosomes (YACs). YACs are linear DNA molecules that contain all the chromosomal structures required for normal replication and segregation during yeast cell division. BACs, which replicate in E. coli, are de¬ rived from circular plasmids that normally replicate long regions of DNA and are maintained at the level of approximately one copy per cell (properties sim¬ ilar to those of actual chromosomes).

Ligase Joins Two DNA Segments. A DNA segment to be cloned is often obtained through the action of restriction endonucleases. Most restriction en¬ zymes cleave DNA to yield sticky ends (Section 3-4A). Therefore, as Janet Mertz and Ron Davis first demonstrated in 1972, a restriction fragment can be inserted into a cut made in a cloning vector by the same restriction enzyme (Fig. 3-24). The complementary ends of the two DNAs form base pairs (an¬ neal) and the sugar-phosphate backbones are covalently ligated, or spliced together, through the action of an enzyme named DNA ligase. (A ligase pro¬ duced by a bacteriophage can also join blunt-ended restriction fragments.) A great advantage of using a restriction enzyme to construct a recombinant DNA molecule is that the DNA insert can later be precisely excised from the cloned vector by cleaving it with the same restriction enzyme.

Selection Detects the Presence of a Cloned DNA. The expression of a chimeric plasmid in a bacterial host was first demonstrated in 1973 by Herbert Boyer and Stanley Cohen. A host bacterium can take up a plasmid when the two are mixed together, but the vector becomes permanently established in its bacterial host (transformation) with an efficiency of only —0.1%. However, a single transformed cell can multiply without limit, producing large quantities of recombinant DNA. Bacterial cells are typically plated on a semisolid growth medium at a low enough density that discrete colonies, each arising from a single cell, are visible.

It is essential to select only those host organisms that have been transformed and that contain a properly constructed vector. In the case of plasmid transfor¬ mation, selection can be accomplished through the use of antibiotics and/or chromogenic (color-producing) substances. For example, the lacZ gene in the pUC18 plasmid (see Fig. 3-22) encodes the enzyme (3-galactosidase, which cleaves the colorless compound X-gal to a blue product:

Cells of E. coli that have been transformed by an unmodified pUC18 plasmid form blue colonies. However, if the plasmid contains a foreign DNA insert in its polylinker region, the colonies are colorless because the insert interrupts the protein-coding sequence of the lacZ gene and no functional (3-galactosidase is produced. Bacteria that have failed to take up any plasmid are also colorless due to the absence of (3-galactosidase, but these cells can be excluded by adding the antibiotic ampicillin to the growth medium (the plasmid includes the gene amp R , which confers ampicillin resistance). Thus, successfully transformed cells form colorless colonies in the presence of ampicillin. Genes such as amp R are known as selectable markers.

Genetically engineered bacteriophage X vectors contain restriction sites that flank the dispensable central third of the phage genome. This

segment can be replaced by foreign DNA, but the chimeric DNA is packaged in phage particles only if its length is from 75 to 105% of the 48.5-kb wild-type X genome (Fig. 3-25). Consequently, X phage vectors that have failed to acquire a foreign DNA insert are unable to propagate because they are too short toform infectious phage particles. Of course, the production of infectious phage particles results not in a growing bacterial colony but in a plaque, a region of lysed bacterial cells, on a culture plate containing a “lawn” of the host bacte¬ ria. The recombinant DNA—now much amplified—can be recovered from the phage particles in the plaque.

B DNA Libraries Are Collections of Cloned DNA

In order to clone a particular DNA fragment, it must first be obtained in rel¬ atively pure form. The magnitude of this task can be appreciated by consid¬ ering that, for example, a 1-kb fragment of human DNA represents only 0.00003% of the 3 billion-bp human genome. Of course, identifying a par¬ ticular DNA fragment requires knowing something about its nucleotide se¬ quence or its protein product. In practice, it is usually more difficult to identify a particular DNA fragment from an organism and then clone it than it is to clone all the organisms DNA that might contain the DNA of interest and then identify the clones containing the desired sequence.

A Genomic Library Includes All of an Organism’s DNA. The cloned set of all DNA fragments from a particular organism is known as its genomic library. Genomic libraries are generated by a procedure known as shotgun cloning. The chromosomal DNA of the organism is isolated, cleaved to frag¬ ments of cloneable size, and inserted into a cloning vector. The DNA is usu¬ ally fragmented by partial rather than exhaustive restriction digestion so that the genomic library contains

intact representatives of all the organisms genes, including those that contain restriction sites. DNA in solution can also be me¬ chanically fragmented (sheared) by rapid stirring. Given the large size of the genome relative to a gene, the shotgun cloning method is subject to the laws of probability. The number of randomly gener¬ ated fragments that must be cloned to ensure a high probability that a desired sequence is represented at least once in the genomic library is calculated as follows:

The probability Pthat a set of clones contains a fragment that con¬ stitutes a fraction f, in bp, of the organisms genome is

P=l-(\-f) N [3-1]

Consequently,

7V=log(l -P)/log(l -f) [3-2]

Thus, in order for P to equal 0.99 for fragments averaging 10 kb in length, N = 2162 for the 4600-kb E. coli chromosome (see Sample Calculation 3-1). The use of BAC- or YAC-based genomic libraries with their large fragment lengths therefore greatly reduces the effort necessary to obtain a given gene seg¬ ment from a large genome. After a BAC- or YAC-based clone containing the desired DNA has been identified (see below), its large DNA insert can be fur¬ ther fragmented and cloned again (subcloned) to isolate the target DNA.

A cDNA Library Represents Expressed Genes. A different type of DNA library contains only the expressed sequences from a particular cell type. Such a cDNA library is constructed by isolating all the cells mRNAs and then copying them to DNA using a specialized type of DNA polymerase

known as reverse transcriptase because it synthesizes DNA using RNA templates (Box 25-2). The complementary DNA (cDNA) molecules are then inserted into cloning vectors to form a cDNA library. A cDNA library can also be used to construct a DNA microarray (DNA chip), in which each different cDNA is immobilized at a specific position on a slide. A DNA chip can be used for detecting the presence of mRNA in a biological sample (the mRNA, if pres¬ ent, will bind to its complementary cDNA; Section 14-4C).

A Library Is Screened for the Gene of Interest. Once the requisite number of clones is obtained, the genomic library must be screened for the presence of the desired gene. This can be done by a process known as colony or in situ hybridization (Latin: in situ. , in position; Fig. 3-26). The cloned yeast colonies, bacterial colonies, or phage plaques to be tested are transferred, by replica plating, from a master plate to a nitrocellulose filter (replica plating is also used to transfer colonies to plates containing different growth media). Next, the filter is treated with NaOH, which lyses the cells or phages and sep¬ arates the DNA into single strands, which preferentially bind to the nitrocel¬ lulose. The filter is then dried to fix the DNA in place and incubated with a labeled probe. The probe is a short segment of DNA or RNA whose sequence is complementary to a portion of the DNA of interest. After washing away unbound probe, the presence of the probe on the nitrocellulose is detected by a technique appropriate for the label used (e.g., exposure to X-ray film for a radioactive probe, a process known as autoradiography, or illumination with an appropriate wavelength for a fluorescent probe). Only those colonies or plaques containing the desired gene bind the probe and are thereby detected. The corresponding clones can then be retrieved from the master plate. Using this technique, a human genomic library of ~ 1 million clones can be readily screened for the presence of one particular DNA segment.

Choosing a probe for a gene whose sequence is not known requires some artistry. The corresponding mRNA can be used as a probe if it is available in sufficient quantities to be isolated. Alternatively, if the amino acid sequence of the protein encoded by the gene is known, the probe may be a mixture of the various synthetic oligonucleotides that are complementary to a segment of the genes inferred base sequence. Several disease-related genes have been iso¬ lated using probes specific for nearby markers, such as repeated DNA se¬ quences, that were already known to be genetically linked to the disease genes.

C DNA Is Amplified by the Polymerase Chain Reaction

Although molecular cloning techniques are indispensable to modern biochem¬ ical research, the polymerase chain reaction (PCR) is often a faster and more convenient method for amplifying a specific DNA. Segments of up to 6 kb can be amplified by this technique, which was devised by Kary Mullis in 1985. In PCR, a DNA sample is separated into single strands and incubated with DNA polymerase, dNTPs, and two oligonucleotide primers whose sequences flank the DNA segment of interest. The primers direct the DNA polymerase to synthesize com¬ plementary strands of the target DNA (Fig. 3-27). Multiple cycles of this process, each doubling the amount of the target DNA, geometrically amplify the DNA starting from as little as a single gene copy. In each cycle, the two strands of the duplex DNA are separated by heating, then the reaction mixture is cooled to allow the primers to anneal to their complementary segments on the DNA. Next, the DNA polymerase directs the synthesis of the complementary strands. The use of a heat-stable DNA polymerase, such as Taq polymerase isolated from Thermus aquaticus, a bacterium that thrives at 75°C, eliminates the need to add fresh enzyme after each round of heating (heat inactivates most en¬ zymes). Hence, in the presence of sufficient quantities of primers and dNTPs, PCR is carried out simply by cyclically varying the temperature.

Twenty cycles of PCR increase the amount of the target sequence around a millionfold (~2 20) with high specificity. Indeed, PCR can amplify a target DNA present only once in a sample of 10 5 cells, so this method can be used without prior DNA purification. The amplified DNA can then be sequenced or cloned.

PCR amplification has become an indispensable tool. Clinically, it is used to diagnose infectious diseases and to detect rare pathological events such as muta¬ tions leading to cancer. Forensically, the DNA from a single hair or sperm can be amplified by PCR so that it can be used to identify the donor (Box 3-2). Traditional ABO blood-type analysis requires a coin-sized drop of blood; PCR is effective on pinhead-sized samples of biological fluids. Courts now consider DNA sequences as unambiguous identifiers of individuals, as are fingerprints, because the chance of two individuals sharing extended sequences of DNA is typically one in a million or more. In a few cases, PCR has dramatically restored justice to con¬ victs who were released from prison on the basis of PCR results that proved their innocence—even many years after the crime-scene evidence had been collected.

Recombinant DNA Technology Has Numerous Practical Applications

The ability to manipulate DNA sequences allows genes to be altered and ex¬ pressed in order to obtain proteins with improved functional properties or to correct genetic defects.

Cloned Genes Can Be Expressed. The production of large quantities of scarce or novel proteins is relatively straightforward only for bacterial proteins: A cloned gene must be inserted into an expression vector, a plasmid that contains properly positioned transcriptional and

translational control se¬ quences. The production of a protein of interest may reach 30% of the host's total cellular protein. Such genetically engineered organisms are called over¬ producers. Bacterial cells often sequester large amounts of useless and possi¬ bly toxic (to the bacterium) protein as insoluble inclusions, which sometimes simplifies the task of purifying the protein.

Bacteria can produce eukaryotic proteins only if the recombinant DNA that carries the protein-coding sequence also includes bacterial transcriptional and translational control sequences. Synthesis of eukaryotic proteins in bacte¬ ria also presents other problems. For example, many eukaryotic genes are large and contain stretches of nucleotides (introns) that are transcribed and excised before translation (Section 26-3A); bacteria lack the machinery to excise the introns. In addition, many eukaryotic proteins are posttranslationally modi¬ fied by the addition of carbohydrates or by other reactions. These problems can be overcome by using expression vectors that propagate in eukaryotic hosts, such as yeast or cultured insect or animal cells.

Table 3-5 lists some recombinant proteins produced for medical and agri¬ cultural use. In many cases, purification of these proteins directly from human or animal tissues is unfeasible on ethical or practical grounds. Expression systems per¬ mit large-scale, efficient preparation of the proteins while minimizing the risk of con¬ tamination by viruses or other pathogens from tissue samples.

Site-Directed Mutagenesis Alters a Gene's Nucleotide Sequence. After iso¬ lating a gene, it is possible to modify the nucleotide sequence to alter the amino acid sequence of the encoded protein. Sitedirected

mutagenesis, a technique pio¬ neered by Michael Smith, mimics the natural

process of evolution and allows predictions about the structural and functional roles of particular amino acids in a protein to be rigorously tested in the laboratory.

Synthetic oligonucleotides are required to specifically alter genes through site-directed mutagenesis. An oligonucleotide whose sequence is identical to a portion of the gene of interest except for the desired base changes is used to direct replication of the gene. The oligonucleotide hybridizes to the correspon¬ ding wild-type (naturally occurring) sequence if there are no more than a few mismatched base pairs. Extension of the oligonucleotide, called a primer, by DNA polymerase yields the desired altered gene (Fig. 3-28). The altered gene can then be inserted into an appropriate vector. A mutagenized primer can also be used to generate altered genes by PCR.

Transgenic Organisms Contain Foreign Genes. For many purposes it is prefer¬ able to tailor an intact organism rather than just a protein—true genetic engi¬ neering. Multicellular organisms expressing a gene from another organism are said to be transgenic, and the transplanted foreign gene is called a transgene.

For the change to be permanent, that is, heritable, a transgene must be stably integrated into the organisms germ cells. For mice, this is accomplished by microinjecting cloned DNA encoding the desired altered characteristics into a fertilized egg and implanting it into the uterus of a foster mother. A wellknown example of a transgenic mouse contains extra copies of a growth hor¬ mone gene (Fig. 3-29).

Transgenic farm animals have also been developed. Ideally, the genes of such animals could be tailored to allow the animals to grow faster on less food or to be resistant to particular diseases. Some transgenic farm

animals have been engineered to secrete medically useful proteins into their milk. Harvesting such a substance from milk is much more cost-effective than producing the same substance in bacterial cultures.

One of the most successful transgenic organisms is corn (maize) that has been genetically modified to produce a protein that is toxic to plant-eating in¬ sects (but harmless to vertebrates). The toxin is synthesized by the soil mi¬ crobe Bacillus thuringiensis. The toxin gene has been cloned into corn in order to confer protection against the European corn borer, a commercially signifi¬ cant pest that spends much of its life cycle inside the corn plant, where it is largely inaccessible to chemical insecticides. The use of "Bt corn," which is now widely planted in the United States, has greatly reduced the need for such toxic substances.

Transgenic plants have also been engineered for better nutrition. For ex¬ ample, researchers have developed a strain of rice with foreign genes that en¬ code enzymes necessary to synthesize P-carotene (an orange pigment that is the precursor of vitamin A) and a gene for the iron-storage protein ferritin. The genetically modified rice, which is named "golden rice" (Fig. 3-30), should help alleviate vitamin A deficiencies (which afflict some 400 million people) and iron deficiencies (an estimated 30% of the worlds population suf¬ fers from iron deficiency). Other transgenic plants include freeze-tolerant strawberries, slow-ripening tomatoes, and rapidly maturing fruit trees.

There is presently a widely held popular suspicion, particularly in Europe, that genetically modified or "GM" foods will somehow be harmful. However, extensive research, as well as considerable consumer experience, has failed to reveal any deleterious effects caused by GM foods (see Box 3-3).

Transgenic organisms have greatly enhanced our understanding of gene expression. Animals that have been engineered to contain a defective gene or hat lack a gene entirely (a so-called gene knockout) also serve as experimen¬ tal models for human diseases.

Genetic Defects Can Be Corrected. Gene therapy is the transfer of new genetic material to the cells of an individual in order to produce a thera¬ peutic effect. Although the potential benefits of this as yet rudimentary tech¬ nology are enormous, there are numerous practical obstacles to overcome. For example, the retroviral vectors (RNA-containing viruses) commonly used to directly introduce genes into humans can provoke a fatal immune response.

The first documented success of gene therapy in humans occurred in children with a form of severe combined immunodeficiency disease (SCID) known as SCID-X1, which without treatment would have required their isolation in a sterile environment to prevent fatal infection. SCID-X1 is caused by a defect in the gene encoding yc cytokine receptor, whose ac¬ tion is essential for proper immune system function. Bone marrow cells (the precursors of white blood cells) were removed from the bodies of SCID-X1 victims, incubated with a vector containing a normal yc cytokine receptor gene, and returned to their bodies. The transgenic bone marrow cells re¬ stored immune system function. However, because the viral vector integrates into the genome at random, the location of the transgene may affect the ex¬ pression of other genes, triggering cancer. At least two children have devel¬ oped leukemia (a white blood cell cancer) as a result of gene therapy for SCID-X1.

Other diseases that have been successfully treated by gene therapy are Leber’s congenital amaurosis, a rare form of blindness, X-linked adrenoleukodystrophy, in which a defect in a membrane transport

protein leads to brain damage, and (3-thalassemia, a type of severe anemia.